Dept. of Math. University of Oslo
Statistical Research Report No. 11
ISSN 0806–3842 December 2005

A robust conflict measure of internal inconsistencies in Bayesian hierarchical models

FREDRIK A. DAHL, Health Services Research Unit, Akershus University Hospital, Norway, and Department of Mathematics, University of Oslo, Norway

JØRUND GÅSEMYR, Department of Mathematics, University of Oslo, Norway

BENT NATVIG, Department of Mathematics, University of Oslo, Norway

Abstract. O'Hagan (2003) introduces some tools for criticism of Bayesian hierarchical models that can be applied at each node of the model, with a view to diagnosing problems of model fit at any point in the model structure. His method relies on computing the posterior median of a conflict index, typically through MCMC simulations. We investigate a Gaussian model of two-way analysis of variance, and show that O'Hagan's approach gives unreliable false warning probabilities. We extend and refine the method, especially avoiding double use of data by a data splitting approach, accompanied by theoretical justifications from a non trivial special case. Through extensive numerical experiments we show that our method detects model misspecification about as well as O'Hagan's method does, while retaining the desired false warning probability for data generated from the assumed model. This also holds for a Student-t version of the model.

Keywords: Bayesian hierarchical models, conflict measure, double use of data, Markov chain Monte Carlo simulations, model evaluation, two-way analysis of variance.

1 Introduction

Modern computer technology combined with MCMC algorithms has made it possible to analyze complex Bayesian hierarchical models. The resulting popularity of complex models has also increased the need for ways of evaluating such models. In a frequentist setting, this is often done by way of p-values, which quantify how surprising the given data set is, under the assumed model. By construction, a frequentist p-value is pre-experimantally uniformly distributed on the unit interval, where low values are interpreted as surprising.

Several Bayesian p-values have been suggested over the last few decades. The so-called *prior predictive p-value* of Box (1980), measures the degree of surprise of the data, according to some metric of choice, under a probability measure defined by the product of the prior and the likelihood given the model. It therefore differs from a frequentist

p-value through the introduction of the prior distribution. The prior predictive p-value is a natural choice in cases where the prior of a Bayesian model represents our true beliefs about the distribution of our parameters prior to seeing data. Usually, however, we apply quite vague priors that represent general uncertainty about parameters that could in principle be arbitrarily precisely estimated with enough data.

In these cases, sampling under the prior makes little sense, and is not even defined for improper priors. Rubin (1984) therefore introduced *posterior predictive p-values* that relies on sampling hypothetical future replications from the posterior distribution. This construction also allows metrics that evaluate discrepancies between data and parameter values, see Gelman et al. (1996). However, posterior predictive p-values use data twice; both directly through the discrepancy function, and indirectly by sampling from the posterior distribution.

This has been criticized by Dey et al. (1998) and Bayarri and Berger (2000), both coming up with alternative approaches. The former paper introduces a simulation based approach where the posterior distribution given the observed data is compared to a medley of posterior distributions given replicated data sets generated from the prior distribution. Hence, the approach is essentially in accordance with the prior predictive approach. The latter paper suggests two variants; the conditional predictive p-value and the partial posterior predictive p-value, both designed to avoid the double use of data by eliminating the influence of a chosen test statistic on the posterior distribution.

Robins et al (2000) proves that the pre-experimental asymptotic distribution of the posterior predictive p-value is more concentrated around 1/2 than a uniform, as opposed to the two variants of Bayarri and Berger (2000). Hence, as also pointed out by Meng (1994) and Dahl (2005), the posterior predictive p-values tend to be conservative in the sense that extreme values get too low probability. Hjort et al. (2006) analyzes this in depth, and designs a double simulation scheme that alleviates the problem. This scheme can be thought of as essentially treating the posterior predictive p-value as a test statistic in itself, and using it in an extensive prior predictive p-value computation.

In model choice problems the task is to choose the best model from a given set of candidates. For Bayesian model choice problems, Bayes factors, see Kass and Raftery (1995) provide a useful methodology. Information criteria give a different approach to model choice based on weighing model fit against the number of free parameters. The Bayesian information criterion (BIC) was defined by Schwartz (1978) and more recently analyzed by Clyde and George (2004). A different information criterion that is used for Bayesian models is the so-called divergence information criteron (DIC), see Spiegelhalter et al. (2002). Although model evaluation and model choice are related, these tasks are different, and model choice methods cannot readily be applied for the purpose of model evaluation.

The variants of Berger and Bayarri (2000) work well in some simple cases, but it seems difficult to use this method to criticize arbitrary aspects of Bayesian hierarchical models. Dey et al (1998) introduces tools for evaluating different parts of such models. In the same spirit, we extend and refine in this paper a tool suggested by O'Hagan

(2003) for evaluating internal inconsistencies of a model, through analysis of what he calls *information contributions*. This is a flexible tool that can be used at any node in the model network. However, in the present paper, we restrict our attention to location parameters. Under suitable conditions, our conflict evaluation for a given node will pre-experimentally be a squared normal variable. Our main hypothesis is that this is close to be true for a larger class of models. This gives a surprise index which is similar to a frequentist p-value. This does not mean that we advocate basing the model building process on a formal hypothesis testing scheme, which would also involve the problem of simultaneous testing of several dependent hypothesis. Rather, we envisage an informal procedure, where the conflict analysis suggests points in the model structure that might be problematic. However, without reasonable control over the pre-experimental distribution of the conflicts in the model, it would be difficult to use this tool in practice without a computationally demanding empirical normalization.

The paper is layed out as follows: In Section 2 we explain the original idea of O'Hagan (2003) in the setting of a Gaussian hierarchical model, followed by our modifications of the method. Our modifications include the splitting of data, in order to avoid double use of it, and this is discussed further in Section 3. Section 4 gives some theoretical results in a special case of our model. In Section 5 we give results from our massive simulation experiments, and Section 6 concludes the article. In the appendix we give the proofs of the theoretical results in Section 4.

2 Measuring conflict

O'Hagan (2003) introduces some tools for model criticism that can be applied at each node of a complex hierarchical or graphical model, with a view to diagnosing problems of model fit at any point in the model structure. In general, the model can be supposed to be expressed as a directed acyclic graph. To compare two unimodal densities/likelihoods he suggests the following procedure. First normalize both densities to have unit maximum height. The height of both curves at their point of intersection is denoted by z. Then the suggested conflict measure is $c^1 = -2 \ln z$. In the present paper we consider, as O'Hagan (2003), the simple hierarchical model for several normal samples (one-way analysis of variance) to clarify what we see as problematic aspects of his approach. Observations y_{ij} for $i = 1, ..., k$ and $j = 1, ..., n_i$ are available. The model has the form:

$$\begin{aligned} y_{ij}|\boldsymbol{\lambda}, \sigma^2 &\sim^{ind} N(\lambda_i, \sigma^2), i = 1, ..., k; j = 1, ..., n_i \\ \lambda_i|\mu, \tau^2 &\sim^{ind} N(\mu, \tau^2), i = 1, ..., k, \end{aligned} \tag{1}$$

and is completed by a prior distribution for σ^2, τ^2 and μ.

In the model (1), consider the node for parameter λ_i. In addition to its parents μ and τ^2, it is linked to its child nodes $y_{i1}, ..., y_{in_i}$. The full conditional distribution of λ_i is

given by:

$$p(\lambda_i|\boldsymbol{y}, \boldsymbol{\lambda}_{-i}, \sigma^2, \tau^2, \mu) \propto p(\lambda_i|\mu, \tau^2) \prod_{j=1}^{n_i} p(y_{ij}|\lambda_i, \sigma^2) \tag{2}$$

This shows how each of the $n_i + 1$ distributions can be considered as a source of information about λ_i. When we are considering the possibility of conflict at the λ_i node, we must consider each of these contributing distributions as functions of λ_i. In the present model, contrasting the information about λ_i from the parent nodes with that from the child nodes, the conflict measure simplifies to:

$$c^1_{\lambda_i} = (\mu - \bar{y}_i)^2/(\tau + \sigma/\sqrt{n_i})^2, \tag{3}$$

where $\bar{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$, noting that the last n_i factors of (2) can be written as $p(\bar{y}_i|\lambda_i, \sigma^2)$.

When the parameters σ^2, τ^2 and μ are given by prior distributions, O'Hagan (2003) suggests using MCMC to estimate the quantity

$$c^{1,\mathbf{y},\text{med}}_{\lambda_i} = M^{\sigma^2,\tau^2,\mu|\mathbf{y}}(c^1_{\lambda_i}), \tag{4}$$

where M denotes the median under the posterior distribution of σ^2, τ^2 and μ. He claims that a value less than 1 should be thought of as indicative of no conflict, whereas values of 4 or more would certainly indicate a conflict to be taken seriously.

A first problem with (3) is the somewhat odd looking denominator. A more natural choice of normalization seems to be

$$c^2_{\lambda_i} = (\mu - \bar{y}_i)^2/(\tau^2 + \sigma^2/n_i) \tag{5}$$

In a simplified situation where σ^2, τ^2 and μ are given numbers, $c^2_{\lambda_i}$ is χ^2_1 distributed pre-experimentally. Hence, in this case we can argue that values of $c^2_{\lambda_i}$ exceeding 4 do indicate a serious conflict. A second problem with the O'Hagan (2003) approach is that data are used twice, first in the computation of the posterior distribution in (4), and then in the evaluation of the conflict measure. One way to avoid this is to split the data in one part $\boldsymbol{y}^1$ used to obtain a posterior distribution for the parameters τ^2, μ of the parent nodes information contribution, and another part $\boldsymbol{y}^2$ used to obtain a posterior distribution for the parameter σ^2 of the child nodes information contribution. In the evaluation of the conflict, we use a data vector $\boldsymbol{y}^2_i$, defined as the components of $\boldsymbol{y}_i$ present in $\boldsymbol{y}^2$.

A third problem concerns the way in which the use of the posterior distributions of the nuisance parameters σ^2, τ^2 and μ affect the level of conflict. It is not at all obvious that the median construction (4) normalizes the conflict in a stable and sensible way. We suggest as an alternative to construct two distributions g_1 and g_2 representing the two different information sources for λ_i, $N(\mu, \tau^2)$ and $N(\bar{y}^2_i, \sigma^2/n_i)$, integrated over the posterior distributions of τ^2, μ given $\boldsymbol{y}^1$ respectively σ^2 given $\boldsymbol{y}^2$. This explains the

abbreviation ipd (integrated posterior distributions) in the following conflict measure, analogous to (5), between g_1 and g_2:

$$c_{\lambda_i}^{2,\boldsymbol{y}^1,\boldsymbol{y}^2,\text{ipd}} = (E^{g_1}(\lambda_i) - E^{g_2}(\lambda_i))^2/(\text{var}^{g_1}(\lambda_i) + \text{var}^{g_2}(\lambda_i)) \tag{6}$$

By a conditional expectation and variance argument, this simplifies to

$$c_{\lambda_i}^{2,\boldsymbol{y}^1,\boldsymbol{y}^2,\text{ipd}} = \frac{(E(\mu|\boldsymbol{y}^1) - \bar{y}_i^2)^2}{E(\tau^2|\boldsymbol{y}^1) + var(\mu|\boldsymbol{y}^1) + E(\sigma^2|\boldsymbol{y}^2)/n_i} \tag{7}$$

Note the additional denominator term $var(\mu|\boldsymbol{y}^1)$ of (7) as opposed to (5).

The ipd construction can be generalized to conflicts concerning arbitrary nodes in the hierarchical network. When σ^2, τ^2 and μ are fixed, the posterior distributions are degenerate. Then (7) coincides with (5) and is hence suitably normalized. However, when these parameters are random, the variance terms in the denominator of $c_{\lambda_i}^{2,\boldsymbol{y},\text{med}}$ capture only part of the pre-experimental variability of the numerator, while the integration over the posterior distributions ensures that the variance terms in the denominator of (7) approximately reflect the different sources of pre-experimental variability of the numerator.

The two basic conflict measures (3) and (5), the various data splittings, and the possibility to choose between the median and the ipd approach, give a large number of possibilities for assessing conflict. In this paper we investigate these possibilities through MCMC simulations, both with respect to a false warning probability or significance level, and with respect to calibrated detection probabilities. We consider those methods that hit reasonably close to an intended false warning probability as greatly preferable, since these methods may make a computationally costly empirical normalization step unnecessary. Among these methods, we prefer those that have the highest detection probability.

3 Data splitting approaches

The consequences of double use of data can be impossible to assess. This motivates the introduction of different data splitting approaches, designed to avoid this.

Visualize the model (1) with the nodes of the parameters σ^2, τ^2 and μ in the first row, the nodes of $\lambda_i, i = 1, ..., k$, in the second row, and the nodes of the transposed of $\boldsymbol{y}_i, i = 1, ..., k$ as columns in the third row. A horizontal splitting of the data $\boldsymbol{y}$ would be achieved by letting

$$\begin{array}{ll} \boldsymbol{y}^1 & = (y_{11}, ..., y_{1m_1}, ..., y_{k1}, ..., y_{km_k}) \\ \boldsymbol{y}^2 & = (y_{1m_1+1}, ..., y_{1n_1}, ..., y_{km_k+1}, ..., y_{kn_k}) \end{array}, \tag{8}$$

where $1 \leq m_i < n_i$ for $i = 1, ..., k$. Let $\boldsymbol{y}_i^2 = (y_{im_i+1}, ..., y_{in_i}), i = 1, ...k$. Furthermore, let $c_{\lambda_i}((\tau^2, \mu); (\sigma^2, \boldsymbol{y}_i))$ be any of the two conflict measures $c_{\lambda_i}^1$ and $c_{\lambda_i}^2$ given by (3) and

(5). To avoid the double use of data the approach of (4) could be replaced by

$$c_{\lambda_i}^{\boldsymbol{y}^1,\boldsymbol{y}^2,\text{med}} = M^{(\tau^2,\mu|\boldsymbol{y}^1)\times(\sigma^2|\boldsymbol{y}^2)}(c_{\lambda_i}((\tau^2,\mu);(\sigma^2,\boldsymbol{y}_i^2))) \tag{9}$$

By running a suitable MCMC algorithm twice, to obtain posterior samples of the parameters σ^2, τ^2, and μ given $\boldsymbol{y}^1$ and $\boldsymbol{y}^2$, respectively, we could estimate all k conflicts $c_{\lambda_i}^{\boldsymbol{y}^1,\boldsymbol{y}^2,\text{med}}, i = 1, ..., k$. Note also that when using (8), $\boldsymbol{y}^1$ and $\boldsymbol{y}^2$ can be interchanged, and the corresponding quantity be estimated from the same two posterior samples of σ^2, τ^2, and μ. If, for n_i even, $m_i = n_i/2, i = 1, ..., k$, equal weights should be allocated to these two parallel estimates.

A vertical splitting avoiding the double use of data would be achieved by letting, for $1 \leq l < k$

$$\begin{aligned} \boldsymbol{y}^1 &= (y_{11}, ..., y_{1n_1}, ..., y_{l1}, ..., y_{ln_l}) \\ \boldsymbol{y}^2 &= (y_{l+11}, ..., y_{l+1n_{l+1}}, ..., y_{k1}, ..., y_{kn_k}) \end{aligned} \tag{10}$$

Let $\boldsymbol{y}_i^2 = (y_{i1}, ..., y_{in_i}), i = l+1, ..., k$. By applying this splitting in (9), we can estimate from two posterior samples the conflicts $c_{\lambda_i}^{\boldsymbol{y}^1,\boldsymbol{y}^2,\text{med}}, i = l+1, ..., k$. The remaining conflicts $c_{\lambda_i}^{\boldsymbol{y}^1,\boldsymbol{y}^2,\text{med}}, i = 1, ..., l$ are estimated by interchanging $\boldsymbol{y}^1$ and $\boldsymbol{y}^2$. Assume we are especially interested in a possible conflict at a specific λ node, and that we want maximum data to arrive at the posterior distribution of the parameters μ and τ^2. We then denote this node by λ_k, and choose $l = k - 1$.

4 Theoretical comparisons

We start this section by giving some needed notation. In our simulation experiments, which we will return to in Section 5, we will compare the O'Hagan (2003) conflict measure $c_{\lambda_k}^1$ given by (3), in the following denoted by A_1, with $c_{\lambda_k}^2$ given by (5) (A_2). Secondly, we will compare his approach of evaluating the median posterior conflict, as in (4) (B_1), with the ipd approach presented in (7) (B_2). We will also compare the non data splitting approach of O'Hagan (2003) (C_1) with two horizontal splitting schemes. The first one (C_2) is based on (8), while the second (C_3) is also based on interchanging $\boldsymbol{y}^1$ and $\boldsymbol{y}^2$. These are also compared with two vertical splitting approaches, based on (10), with $l = k/2$ (C_4), and $l = k - 1$ (C_5), respectively. Note that C_4 is only defined for even k. Altogether we have $2 \cdot 2 \cdot 5 = 20$ combinations.

What we call the null model is the model (1) with the nuisance parameters set at their prior expected values, $\sigma^2 = \sigma_0^2$, $\tau^2 = \tau_0^2$, and $\mu = \mu_0$. Denote the actual significance level by α, using a nominal warning level of $\chi^2_{1,0.95} \approx 3.85$ ($\chi^2_{2,0.95}/2 \approx 3.00$ for C_3). β is the detection probability of an alternative fixed λ_k, substantially different from μ_0, when the conflict measure has been calibrated to give a significance level of 0.05.

We now compare some of the approaches theoretically.

First note that since a sum of variances is less than the corresponding square of the summed standard deviations, we obviously have

PROPOSITION 1 *For any splitting*

$$c_{\lambda_k}^{1,\boldsymbol{y}^1,\boldsymbol{y}^2,med} < c_{\lambda_k}^{2,\boldsymbol{y}^1,\boldsymbol{y}^2,med} \text{ and } c_{\lambda_k}^{1,\boldsymbol{y}^1,\boldsymbol{y}^2,ipd} < c_{\lambda_k}^{2,\boldsymbol{y}^1,\boldsymbol{y}^2,ipd} \tag{11}$$

Consequently

$$\alpha_{A_1,B_i,C_j} < \alpha_{A_2,B_i,C_j} \text{ for } i = 1,2; j = 1,...,5. \tag{12}$$

In the following we focus mainly on the case A_2 of the first factor, and compare the cases B_1 and B_2 of the second factor, and the cases C_1, C_2, C_4 and C_5 of the third factor, leaving out the more complex case C_3. In order to make a theoretical analysis tractable, we choose the improper prior 1 for μ, whereas we start out with arbitrary prior distributions for σ^2 and τ^2. In the comparison of the different splittings and in the comparison of the median and the ipd approaches we make the further simplifying assumption that σ^2 and τ^2 are fixed. We fix $n_i = n$, for n even, and, in the case of the splitting (8), $m_i = n/2$ for each $i = 1, \ldots, k$. Proofs of the main results are given in the appendix.

Our first aim is to analyse how close the ipd conflict $c_{\lambda_k}^{2,\boldsymbol{y}^1,\boldsymbol{y}^2,\text{ipd}}$ given by (7) is to being χ_1^2-distributed pre-experimentally under the null model. In order to handle this problem we denote by $\boldsymbol{Y}$ a random replicate of $\boldsymbol{y}$. Also, we denote by E_0 respectively var_0 the expectation and variance under the null model distribution.

THEOREM 1 *Based on the splitting (10) we have*

$$c_{\lambda_k}^{2,\boldsymbol{Y}^1,\boldsymbol{Y}^2,ipd} = (\bar{Y}^1 - \bar{Y}_k^2)^2 / \hat{var}(\bar{Y}^1 - \bar{Y}_k^2), \tag{13}$$

where, for given data $\boldsymbol{y}^1, \boldsymbol{y}^2$, $\hat{var}(\bar{y}^1 - \bar{y}_k^2)$ *is the estimate*

$$(1/n)[E(\sigma^2|\boldsymbol{y}^2) + (1/l)E(\sigma^2|\boldsymbol{y}^1)] + ((l+1)/l)E(\tau^2|\boldsymbol{y}^1) \tag{14}$$

of

$$var_0(\bar{Y}^1 - \bar{Y}_k^2) = ((l+1)/l)((1/n)\sigma_0^2 + \tau_0^2). \tag{15}$$

The denominator of (13) is independent of $\bar{Y}^1 - \bar{Y}_k^2$.

This theorem justifies our assertion in Section 2 that with the ipd approach, the variance terms in the denominator of (7) approximately reflect the different sources of variability of the numerator.

However, the uncertainty in the posterior distributions of the variance parameters σ^2 and τ^2 may give a slight exaggeration of the conflict. To see this, note that it follows

from Theorem 1 that under the null model we may write $c_{\lambda_k}^{2,\boldsymbol{Y}^1,\boldsymbol{Y}^2,\text{ipd}} = XZ$, where X and Z are independent, and

$X = (\bar{Y}^1 - \bar{Y}_k^2)/\text{var}_0(\bar{Y}^1 - \bar{Y}_k^2)$ is χ_1^2-distributed, and

$Z = \text{var}_0(\bar{Y}^1 - \bar{Y}_k^2)/\hat{\text{var}}(\bar{Y}^1 - \bar{Y}_k^2)$.

Due to Theorem 1 we assume $E_0(Z) \approx 1$. Hence, we obtain

$\text{var}_0(c_{\lambda_k}^{2,\boldsymbol{Y}^1,\boldsymbol{Y}^2,\text{ipd}}) = \text{var}_0(XZ) = (E_0(Z))^2\text{var}_0(X) + E_0(X^2)\text{var}_0(Z) \approx \text{var}_0(X) + E_0(X^2)\text{var}_0(Z) \geq \text{var}_0(X) = \text{var}(\chi_1^2)$.

This makes it reasonable to believe that the distribution of $c_{\lambda_k}^{2,\boldsymbol{Y}^1,\boldsymbol{Y}^2,\text{ipd}}$ has a heavier tail than the χ_1^2-distribution, and hence that

$\alpha_{A_2,B_2,C_j} >$ or ≈ 0.05 for $j = 4, 5$.

When the variance parameters are fixed, but still with an improper prior for μ, the denominator of (13) reduces to (15). Hence, the following corollary is a direct consequence of Theorem 1.

COROLLARY 1 *Let $\sigma^2 = \sigma_0^2$ and $\tau^2 = \tau_0^2$ be fixed and known. Then, with the splitting (10), the conflict $c_{\lambda_k}^{2,\boldsymbol{Y}^1,\boldsymbol{Y}^2,ipd}$ is χ_1^2- distributed under the null model, and $\alpha_{A_2,B_2,C_j} = 0.05$ for $j = 4, 5$.*

Equations (13) and (14) hold for the splitting (8) as well, with n replaced by $n/2$ and l replaced by k. With k in place of l, these equalities are true also without any splitting, using $(\bar{Y} - \bar{Y}_k)^2$ in the numerator. However, the variance expression (15) is not valid in these cases, and consequently the conflicts are not χ_1^2- distributed. Instead, we have the following proposition:

PROPOSITION 2 *Suppose $\sigma^2 = \sigma_0^2$ and $\tau^2 = \tau_0^2$ are fixed and known. With no splitting we have*

$c_{\lambda_k}^{2,\boldsymbol{Y},ipd} \sim ((k-1)/(k+1))\chi_1^2$.

With the splitting (8) we have

$c_{\lambda_k}^{2,\boldsymbol{Y}^1,\boldsymbol{Y}^2,ipd} \sim [(k-1)\tau_0^2 + ((k+1)/(n/2))\sigma_0^2]/[(k+1)(\tau_0^2 + (1/(n/2))\sigma_0^2)]\chi_1^2$.

Consequently, $\alpha_{A_2,B_2,C_1} < \alpha_{A_2,B_2,C_2} < 0.05$. These upper and lower bounds for α_{A_2,B_2,C_2} are approached as limits when τ_0^2/σ_0^2 respectively σ_0^2/τ_0^2 approaches 0.

In the calculations leading to this proposition, we used the fact that the numerator in the conflict based on no splitting is

$(\bar{Y} - \bar{Y}_k)^2 = ((k-1)/k)^2(((1/(k-1))(\bar{Y}_1 + \ldots + \bar{Y}_{k-1}) - \bar{Y}_k)^2$.

The right hand side of this equality is $((k-1)/k)^2$ times the numerator in the conflict based on the splitting (10) with $l = k-1$. Since the variance parameters are fixed, it follows that the conflict based on this splitting is proportional to the conflict based on no splitting, and consequently these two conflict measures have identical calibrated detection probabilities, i.e. $\beta_{A_2,B_2,C_1} = \beta_{A_2,B_2,C_5}$.

We now turn to the comparison of the median and the ipd approaches. We have the following proposition

PROPOSITION 3 *Suppose $\sigma^2 = \sigma_0^2$ and $\tau^2 = \tau_0^2$ are fixed and known. We then have for any of the splittings (including no splitting)*

$$c_{\lambda_k}^{2,\boldsymbol{y}^1,\boldsymbol{y}^2,med} > (\bar{y}_k^2 - \bar{y}^1)^2/((1/m_k)\sigma_0^2 + \tau_0^2) >$$

$$(\bar{y}_k^2 - \bar{y}^1)^2/(((l+1)/l))((1/m_k)\sigma_0^2 + \tau_0^2)) = c_{\lambda_k}^{2,\boldsymbol{y}^1,\boldsymbol{y}^2,ipd},$$

where $l = k$ in cases C_1 and C_2, and $m_k = n/2$ in case C_2, whereas $m_k = n$ otherwise. Also, $\alpha_{A_2,B_1,C_j} > \alpha_{A_2,B_2,C_j}$ for $j = 1,2,4,5$. Finally, for any of the splittings, any combination of A and B give identical calibrated detection probabilities, i.e. $\beta_{A_1,B_1,C_j} = \beta_{A_1,B_2,C_j} = \beta_{A_2,B_1,C_j} = \beta_{A_2,B_2,C_j}$ for $j = 1,2,4,5$.

The first of these inequalities shows that the conflict is exaggerated because taking the median exaggerates the numerator, whereas the second inequality demonstrates an additional exaggeration effect arising from a too small variance term in the denominator. This justifies our assertion in Section 2 that for the median conflict, the variance terms in the denominator of (5) capture only part of the variability of the numerator.

5 Simulation experiments

In this section we present the results of some simulation experiments, designed to evaluate false alarm probabilities (α's) and calibrated detection probabilities (β's) for the given model.

We assume that the prior distributions for σ^2, τ^2, and μ are independent. The parameters σ^2 and τ^2 are inverse gamma distributed, both with shape parameter 4, and scale parameters 12 and 3, respectively. The prior distribution for μ is normal with mean 0 and variance $\omega^2 = 9$. Furthermore, we choose $k = 6$, and $n_i = n, i = 1, ..., k$. We run identical experiments with $n = 10$ and $n = 100$.

In the following subsection we present results for the model, with normally distributed $\boldsymbol{\lambda}$. In the next subsection, we analyze a modified version with Student-t distributed $\boldsymbol{\lambda}$, in order to illustrate how departure from normality affects the results.

5.1 Normally distributed $\boldsymbol{\lambda}$

From (1) we arrive at the following likelihood for σ^2, τ^2, and μ

$$
\begin{aligned}
&L(\sigma^2, \tau^2, \mu | y_{ij}, i = 1, ..., k; j = 1, ...n) \\
&= \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} (\frac{1}{\sqrt{2\pi}\sigma})^{nk} e^{-\frac{1}{2\sigma^2}\sum_{i,j}(y_{ij}-\lambda_i)^2} (\frac{1}{\sqrt{2\pi}\tau})^k e^{-\frac{1}{2\tau^2}\sum_i(\lambda_i-\mu)^2} d\lambda_1 ... d\lambda_k \\
&= (\frac{1}{\sqrt{2\pi}\sigma})^{(n-1)k} e^{-\frac{1}{2\sigma^2}\sum_{i,j}(y_{ij}-\bar{y}_i)^2} n^{-k/2} \prod_{i=1}^{k} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma/\sqrt{n}} e^{-\frac{1}{2\sigma^2/n}(\bar{y}_i-\lambda_i)^2} \frac{1}{\sqrt{2\pi}\tau} e^{-\frac{1}{2\tau^2}(\lambda_i-\mu)^2} d\lambda_i \\
&= \prod_{i=1}^{k} (\frac{1}{\sqrt{2\pi}\sigma})^{n-1} \frac{1}{\sqrt{n}} e^{-\frac{1}{2\sigma^2}\sum_{j=1}^{n}(y_{ij}-\bar{y}_i)^2} \frac{1}{\sqrt{2\pi(\sigma^2/n+\tau^2)}} e^{-\frac{1}{2(\sigma^2/n+\tau^2)}(\bar{y}_i-\mu)^2} \qquad (16)
\end{aligned}
$$

From the prior distributions and (16) we generated posterior samples of (σ^2, τ^2, μ) by the Metropolis-Hastings algorithm. We used a random walk version of the algorithm, with simultaneous steps in each direction. The steps are constructed as mixtures of centered uniform variables with varying size. After a burn-in of 10^6 steps, we generated a sample of 10.000 parameter vectors, sampled 1000 simulation steps apart. This gives close to no serial correlation in the sampled points. We do not claim that this is optimal in any sense, but it runs sufficiently fast for the present application.

Also a total of 10.000 independent data sets were generated in separate files, each containing $kn = 60$ and $kn = 600$ observations. In the null model, the prior expected values are $\sigma_0^2 = 4$, $\tau_0^2 = 1$ and $\mu_0 = 0$. The corresponding 10.000 alternative data sets have $\lambda_k = 3$, i.e., λ_k is located in the tail of its null model distribution $N(0, 1)$.

The C_1 approach requires only one MCMC run, whereas C_2, C_4, C_5 require two MCMC runs each, for a total of seven MCMC runs. These runs must be carried through for 10.000 data sets both from the null model and the alternative model. The total number of MCMC runs is therefore $7 \cdot 2 \cdot 10.000 = 140.000$. In order to test our different conflict measures, we have stored the 140.000 posterior distributions in separate files, each contaning 10.000 parameter triplets (μ, τ^2, σ^2). Due to our elimination of the λ's in the likelihood, the MCMC simulations run very efficiently. Therefore, the main challenge of the process has been the management of the posterior distribution files, rather than computing power.

In Table 1 we give the α and β values estimated from the experiments with $n = 10$. In each cell, the upper number is the estimated value, whereas the lower number is the estimated standard error. Table 2 gives the corresponding results for $n = 100$.

			C_1	C_2	C_3	C_4	C_5
A_1	B_1	α	0	0.007	0.003	0.026	0.024
			0.0001	0.0003	0.0002	0.0010	0.0013
		β	0.70	0.41	0.65	0.64	0.69
			0.0065	0.0038	0.0058	0.0063	0.0152
	B_2	α	0	0.006	0.003	0.023	0.021
			0.0001	0.0003	0.0002	0.0008	0.0014
		β	0.71	0.40	0.65	0.64	0.69
			0.0065	0.0039	0.0058	0.0064	0.0149
A_2	B_1	α	0.026	0.053	0.057	0.104	0.091
			0.0006	0.0006	0.0009	0.0018	0.0033
		β	0.70	0.40	0.65	0.64	0.69
			0.0066	0.0041	0.0057	0.0064	0.0148
	B_2	α	0.008	0.031	0.029	0.051	0.055
			0.0004	0.0004	0.0020	0.0013	0.0028
		β	0.70	0.45	0.64	0.65	0.69
			0.0067	0.0040	0.0056	0.0063	0.0141

Table 1: Simulation results for α and β with n=10. The upper number in a cell is the estimated value, whereas the lower number is the estimated standard error.

			C_1	C_2	C_3	C_4	C_5
A_1	B_1	α	0.003	0.004	0.010	0.065	0.053
			0.0002	0.0003	0.0005	0.0012	0.0024
		β	0.93	0.86	0.93	0.86	0.92
			0.0038	0.0045	0.0032	0.0062	0.0069
	B_2	α	0.002	0.003	0.007	0.054	0.045
			0.0002	0.0002	0.0004	0.0011	0.0021
		β	0.93	0.82	0.93	0.86	0.92
			0.0037	0.0058	0.0032	0.0062	0.0067
A_2	B_1	α	0.021	0.031	0.054	0.121	0.102
			0.0008	0.0005	0.0009	0.0018	0.0028
		β	0.91	0.79	0.91	0.85	0.91
			0.0043	0.0061	0.0043	0.0064	0.0077
	B_2	α	0.004	0.010	0.020	0.056	0.060
			0.0002	0.0004	0.0007	0.0011	0.0024
		β	0.91	0.84	0.92	0.85	0.91
			0.0041	0.0052	0.0045	0.0059	0.0076

Table 2: Simulation results for α and β with n=100. The upper number in a cell is the estimated value, whereas the lower number is the estimated standard error.

We observe from Table 1 and Table 2 that, in accordance with Theorem 1 and Corollary 1, we obtain significance levels quite close to 0.05 when combining our suggested modifications of the approach of O'Hagan (2003), provided we use the vertical split-

tings (10). In fact, we have $\alpha_{A_2,B_2,C_j} \in [0.051, 0.060]$ for $j = 4, 5$. The significance levels exceed 0.05 slightly, as suggested by the discussion following Theorem 1. We also note that replacing A_2 by A_1 results in a substantial drop in significance level for all combinations of the factors B and C, confirming Proposition 1. For any given combination of the factors A and B, such a drop is also observed when replacing C_4, C_5 with the no splitting option C_1, and to a somewhat smaller extent with the horizontal splittings (8), represented by C_2 and C_3. For the combinations $A_2, B_2, C_j, j = 1, 2$, this is in accordance with Corollary 1 and Proposition 2. On the other hand, for any combination of A and C we observe that replacing B_2 with B_1 results in an increase in significance level. This is to be expected from Proposition 3 for the combinations $A_2, C_j, j = 1, 2, 4, 5$. The net effect of combining the conflict reducing factors A_1, C_1 with the conflict increasing factor B_1, which constitutes the original suggestion by O'Hagan (2003), is a significance level dramatically smaller than 0.05. However, for some combinations the downward and upward acting factors cancel, resulting in significance levels fairly close to 0.05. This is the case for the combinations A_2, B_1, C_j with $j = 2, 3$ when $n = 10$, and A_1, B_1, C_j with $j = 4, 5$, as well as A_2, B_1, C_3, when $n = 100$. The combinations $A_1, B_2, C_j, j = 4, 5$, also give acceptable significance levels when $n = 100$, despite the conflict reducing effect of the factor A_1. This is probably due to the fact that this effect is relatively small when $n = 100$, since then σ^2/n is much smaller than τ^2.

Turning to calibrated detection probabilities, we observe that, in accordance with Proposition 3, for any given splitting the detection probabilities are almost the same for all combinations of A and B. The splitting C_2 appears to be somewhat exceptional in this respect. The most important feature influencing the calibrated detection probability seems to be the amount of data used in the estimation of the nuisance parameters for the different splittings. Comparing these gives the ordering C_1, C_5, C_4, C_2, with C_5 almost at the level of C_1, and with C_3 at the level of C_4 for $n = 10$ respectively C_1 and C_5 when $n = 100$.

The combinations $(A_2, B_2, C_j), j = 4, 5$ give false warning probabilities close to the 0.05 significance level. No other combination does this both for $n = 10$ and $n = 100$. Among the vertical splittings, C_5 obtains a calibrated detection probability practically at the level of the no splitting alternative C_1. The symmetric vertical splitting C_4, which uses less data in the estimation of the nuisance parameters, has a somewhat lower detection probability. The relative difference is smaller in the case of abundant data ($n = 100$ compared to $n = 10$). However, the C_4 splitting has the advantage of being able to handle all the 6 possible conflicts at the λ_i-nodes with only 2 MCMC runs. The splitting C_5 needs $2 \cdot 6 = 12$ MCMC runs to evaluate all these conflicts.

5.2 Student-t distributed λ

The practical usefulness of our method will be limited, if it only applies to normal models. We have therefore made an experiment with a non-normal version of our model. There are of course infinitely many ways for a model to be non normal, and

there is no obvious canonical choice. However, the normal distribution has very light tails, and "real data" tend to have a higher probability for extreme values. We have therefore made our experiment with a heavier tailed distribution. A natural choice was the Student-t distribution. We set the degrees of freedom to 3, in order to make the tails as heavy as possible, while still having a finite variance.

Our original model has normal distributions both for λs and for data given λ_i. There is little point in changing the distribution of the data to t-distributions, because the average of these data will be close to normal anyway, due to the central limit theorem. We have therefore chosen to modify the distribution of the λs only, setting $\lambda_i|\mu, \tau^2 \sim T_3$ scaled and located so that $E[\lambda_i|\mu, \tau^2] = \mu$ and $Var[\lambda_i|\mu, \tau^2] = \tau^2$. We somewhat arbitrarily chose $n = 10$ for the experiment. This is not likely to be important, as it made little difference in the normal case.

In our MCMC simulations, we again use the total likelihood directly, and simulate the λ_i parameters together with μ, τ^2, σ^2, using the same Metropolis Hastings algorithm as before. This is rather less efficient than our simulations for the normal model, where we were able to eliminate the λs from the likelihood expression, but still sufficiently fast for our purpose.

Our main hypothesis is that the α level, the false warning probability, is close to 0.05, for our vertical splitting schemes. We have focused on the central splitting (C_4), because this makes it possible to gather data for all $k = 6$ λ_i nodes from the same experiment, due to symmetry of the model.

Following the procedure of our original experiment, we generated 10.000 data sets from the model. The estimated α-level was 0.042, with a standard error of 0.0008. This supports our hypothesis that our method is robust with respect to deviations from normality.

6 Conclusions

We have shown that although the original procedure of O'Hagan (2003) for evaluating conflict is unreliable even in a Gaussian setting, our improvements give a method that can detect problems with a proposed model. Our method is backed up by theoretical computations in a non trivial special case. It is particularly encouraging that our experiments show a false warning probability close to the pre set value of 0.05 for our Gaussian model, and that this appears robust with respect to the normality assumption.

This work has been based on theoretical analysis and experiments with computer generated data sets. A computational approach is in most cases the only way of testing a method's ability to detect deviations from an assumed model, and to evaluate its false warning probability when data is in fact generated from the assumed model. However, the obvious line for future work is to test our method on real data.

Acknowledgement

This work has benefited from the "Evaluation of Bayesian Hierarchical Models" program, supported by the Research Council of Norway, under grant 154911/V30. We are also grateful to Alan Gelfand for his very useful comments on earlier drafts of the manuscript.

References

Bayarri, M. J. & Berger, J. O. (2000). P values for composite null models, *Journal of the American Statistical Association (JASA)* **95**, 1127-1142.

Box, G. E. P. (1980). Sampling & Bayes inference in scientific modeling and robustness, *Journal of the Royal Statistical Society*, Ser. A, **143**, 383-430.

Clyde, M. & George, I. I. (2004). Model uncertainty, *Statistical Science*, 19 (1), 81-94.

Dahl, F. A. (2005). On the conservativeness of posterior predictive p-values, Statistical Research Report Preprint 4, 2005, http://www.math.uio.no/eprint/stat_report/2005/04-05.html (Accepted with minor revision by *Statistics and Probability Letters*).

Dey, D. K., Gelfand, A. E., Swartz, T. B. & Vlachos, P. K. (1998). A simulation-intensive approach for checking hierarchical models. *Test*, **7**, 325-346.

Gelman, A., Meng, X. L. & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statistica Sinica* **6**, 733-807.

Hjort, N. L., Dahl, F. A. & Steinbakk, G. H. (2006). Post-processing posterior predictive p-values, to appear in *Journal of the American Statistical Association*, Theory and Methods.

Kass, R. E. & Raftery, A. E. (1995). Bayes factors, *Journal of the Americal Statistical Association*, **90**, No. 430, 773-795.

Meng, X. L. (1994). Posterior predictive p-values. *The Annals of Statistics*, **22**, 1142-1160.

O'Hagan, A. (2003). HSSS model criticism, in *Highly Structured Stochastic Systems.* Eds: P.J. Green, N.L. Hjort and S. Richardson, Oxford.

Robins, J. M., van der Vaart, A. & Ventura V. (2000). Asymptotic distributions of p values in composite null models (with discussion and rejoiner). *Journal of the Americal Statistical Association*, **95**, 1143-1156.

Schwartz, G. (1978). Estimating the dimension of a model, *Annals of Statistics* **6**, 461-464.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. & van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion and rejoiner). *Journal of the Royal Statistical Society B*, **64**, 583-639.

Appendix

Proof of Theorem 1:

Due to the improper prior for μ we have $\pi(\mu|\boldsymbol{y}^1, \sigma^2, \tau^2) = N(\bar{y}^1, (1/l)((\sigma^2/n) + \tau^2))$, leading to the following simplifications in (7) $E(\mu|\boldsymbol{y}^1) = \bar{y}^1$,

$E(\tau^2|\boldsymbol{y}^1) + \mathrm{var}(\mu|\boldsymbol{y}^1) = E(\tau^2|\boldsymbol{y}^1) + E(\mathrm{var}(\mu|\sigma^2, \tau^2, \boldsymbol{y}^1)) + \mathrm{var}(E(\mu|\sigma^2, \tau^2, \boldsymbol{y}^1)) =$

$E(\tau^2|\boldsymbol{y}^1) + (1/l)(E(\tau^2|\boldsymbol{y}^1) + (1/n)E(\sigma^2|\boldsymbol{y}^1)) = ((l+1)/l)E(\tau^2|\boldsymbol{y}^1) + (1/ln)E(\sigma^2|\boldsymbol{y}^1).$

This proves (14). Moreover, using (16) it can be shown that the posterior distributions of (σ^2, τ^2) given $\boldsymbol{y}^1$ respectively $\boldsymbol{y}^2$ depend on $\boldsymbol{y}^1, \boldsymbol{y}^2$ only through sums of squared differences to the mean, proving the independence assertion. Finally, the variance expression (15) follows from the fact that with the splitting (10), $\boldsymbol{Y}^1$ and $\boldsymbol{Y}^2$ are independent under the null model.

Proof of Proposition 2:

With no splitting we get

$\mathrm{var}_0(\bar{Y} - \bar{Y}_k) = \mathrm{var}_0[((k-1)/k)((\bar{Y}_1 + \ldots + \bar{Y}_{k-1})/(k-1)) - \bar{Y}_k)] =$

$((k-1)/k)^2((\sigma_0^2/n) + \tau_0^2)((1/(k-1)) + 1) = ((k-1)/k)((\sigma_0^2/n) + \tau_0^2).$

On the other hand, for $l = k$ (15) takes the form $((k+1)/k)(\sigma_0^2/n + \tau_0^2)$, and it follows that under no splitting, $c_{\lambda_k}^{2,\boldsymbol{Y},\mathrm{ipd}} \sim ((k-1)/(k+1))\chi_1^2$.

To analyse the splitting (8) we express the data in the form

$Y_{k,j} = \lambda_k + \eta_{k,j}, j = n/2+1, \ldots, n, Y_{i,j} = \lambda_i + \epsilon_{i,j}, i = 1, \ldots, k, j = 1, \ldots, n/2,$

where the variables $\epsilon_{i,j}$ and $\eta_{i,j}$ are independent $\sim N(0, \sigma_0^2)$. With this decomposition, remembering that n is replaced by $n/2$ and l is replaced by k, using (15), we can write the conflict as

$[(\lambda_k - \bar{\lambda}) + (\bar{\eta} - \bar{\epsilon})]^2/[((k+1)/k)(\tau_0^2 + (1/(n/2))\sigma_0^2)] = [((k-1)/k))(\lambda_k - (1/(k-1))(\lambda_1 + \ldots + \lambda_{k-1}) + (\bar{\eta} - \bar{\epsilon})]^2/[((k+1)/k)(\tau_0^2 + (1/(n/2))\sigma_0^2)] =$

$[[((k-1)/k))(\lambda_k - (1/(k-1))(\lambda_1 + \ldots + \lambda_{k-1}) + (\bar{\eta} - \bar{\epsilon})]^2/[((k-1)/k)\tau_0^2 + (2(k+1)/nk)\sigma_0^2]] \times [((k-1)/k)\tau_0^2 + (2(k+1)/nk)\sigma_0^2]/[((k+1)/k)(\tau_0^2 + (1/(n/2))\sigma_0^2)] \sim$

$[(k-1)\tau_0^2 + ((k+1)/(n/2))\sigma_0^2]/[(k+1)(\tau_0^2 + (1/(n/2))\sigma_0^2)]\chi_1^2$. From this the proposition follows.

Proof of Proposition 3:

When $\sigma^2 = \sigma_0^2$ and $\tau^2 = \tau_0^2$ are fixed, the median conflict is $M^{\mu|\boldsymbol{y}^1}((\mu-\bar{y}_k^2)^2/((1/m_k)\sigma_0^2+\tau_0^2))$, where, as before, m_k is either n (no splitting or (10)) or $n/2$ (the splitting (8)). Since μ is the only random quantity in the conflict, and since $E(\mu|\boldsymbol{y}^1) = \bar{y}^1$, the median of $(\mu - \bar{y}_k^2)^2/((1/m_k)\sigma_0^2 + \tau_0^2)$ can be found by solving the equation

$$P(\mu > \bar{y}_k^2 + v|\boldsymbol{y}^1) + P(\mu < \bar{y}_k^2 - v|\boldsymbol{y}^1) = 1/2,$$

or equivalently, assuming without loss of generality that $\bar{y}^1 \geq \bar{y}_k^2$ and letting $v = (\bar{y}^1 - \bar{y}_k^2) + z$,

$$P(\mu > \bar{y}^1 + z|\boldsymbol{y}^1) + P(\mu < \bar{y}_k^2 - (\bar{y}^1 - \bar{y}_k^2) - z|\boldsymbol{y}^1) = 1/2,$$

Clearly, since $P(\mu > \bar{y}^1|\boldsymbol{y}^1) = 1/2$, z must be positive, and the resulting median conflict is

$$(\bar{y}^1 + z - \bar{y}_k^2)^2/((1/m_k)\sigma_0^2 + \tau_0^2).$$

Hence,

$$c_{\lambda_k}^{2,\boldsymbol{y}^1,\boldsymbol{y}^2,\text{med}} > (\bar{y}^1 - \bar{y}_k^2)^2/((1/m_k)\sigma_0^2 + \tau_0^2) >$$

$$(\bar{y}^1 - \bar{y}_k^2)^2/(((l+1)/l))((1/m_k)\sigma_0^2 + \tau_0^2)) = c_{\lambda_k}^{2,\boldsymbol{y}^1,\boldsymbol{y}^2,\text{ipd}}.$$

This covers (10). In the case of (8) or no splitting, $(l+1)/l)$ must be replaced by $(k+1)/k$. This proves the first part, and the corresponding inequalities between the significance levels are immediate. It is seen that z is a deterministic, monotonically decreasing function of $\bar{y}^1 - \bar{y}_k^2$. This implies a deterministic, monotonic relationship between the median and the ipd conflicts $c_{\lambda_k}^{i,\boldsymbol{y}^1,\boldsymbol{y}^2,\text{med}}$ and $c_{\lambda_k}^{i,\boldsymbol{y}^1,\boldsymbol{y}^2,\text{ipd}}$ for $i = 2$ and also for $i = 1$. It follows that the calibrated detection probabilities are identical for a given splitting, i.e. $\beta_{A_1,B_1,C_j} = \beta_{A_1,B_2,C_j} = \beta_{A_2,B_1,C_j} = \beta_{A_2,B_2,C_j}$ for $j = 1, ..., 5$, as asserted.